昆 虫 学 报 *Acta Entomologica Sinica*, February 2010, 53(2): 192 – 201 ISSN 0454-6296

Outgroup selection in tree reconstruction: a case study of the family Halictidae (Hymenoptera: Apoidea)

LUO A-Rong[1,2], ZHANG Yan-Zhou[1], QIAO Hui-Jie[1], SHI Wei-Feng[3], Robert W. MURPHY[4], ZHU Chao-Dong[1,*]

(1. Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China;
2. Graduate University of Chinese Academy of Sciences, Beijing 100049, China;
3. UCD Conway Institute of Biomolecular and Biomedical Sciences, University College Dublin, Dublin 4, Ireland;
4. Department of Natural History, Royal Ontario Museum, 100 Queen's Park, Toronto, ON, M5S 2C6, Canada)

Abstract: An outgroup roots a network to form a tree and/or to infer hypothetical ancestral character states. Usually, multiple taxa of a closely related sister group of the ingroup are selected. To empirically evaluate the choice of outgroup, we implemented three strategies of outgroup selection: a single taxon from the sister group, multiple taxa within the sister group, and multiple taxa from successive sister groups. Subsequently, we evaluated their effects on tree topologies within the family Halictidae (Hymenoptera: Apoidea) incorporating three tree reconstruction methods: maximum likelihood, maximum parsimony and Bayesian inference. The use of multiple taxa within the sister group produced more consistent results than the other two outgroup strategies. The tree topologies were generally consistent with the putative tree topology of Halictidae. Compared with the other two tree reconstruction methods, maximum parsimony produced more consistent results with different outgroup strategies, yet often obtained less resolution.

Key words: Halictidae; homoplasy; monophyly; outgroup; sister group; phylogeny

1 INTRODUCTION

1.1 Importance of outgroup selection

As rigorous methods of tree reconstruction developed, they were widely applied in phylogenetics, molecular ecology, bioconservation, disease control, DNA barcoding and other fields (Cavalli-Sforza and Edwards, 1967; Phillips *et al.*, 2000; Sanderson and Shaffer, 2002). Phylogenies are hypotheses of relationships. The strength of the hypotheses depends on the assumptions of each method of phylogenetic inference. Various assumptions and factors exert effects on tree topologies whether the target organisms include prokaryotes or eukaryotes. For DNA sequence data, the factors can include taxon sampling, alignment methodology, gene selection, gene sequence length, data treatment, optimality criterion, parameter values and others (Smith, 1994; Dalevi *et al.*, 2001; Cameron *et al.*, 2004; Ware *et al.*, 2008). Among these explicit and implicit factors, outgroup selection is usually inadequately considered. Apparently, outgroup selection is often arbitrary or based on obscure relationships between outgroup and ingroup taxa (Lyons-Weiler *et al.*, 1998; Sanderson and Shaffer, 2002; Cameron *et al.*, 2004).

Traditionally, an outgroup serves to root unrooted networks and/or to infer hypothetical ancestral states (Watrous and Wheeler, 1981; Maddison *et al.*, 1984; Wheeler, 1990; Smith, 1994; Lyons-Weiler *et al.*, 1998; Sanderson and Shaffer, 2002). The ingroup should not be studied in isolation. Outgroup selection is critical because topology of the ingroup tree can vary with the choice of outgroup taxa (Milinkovitch and Lyons-Weiler, 1998; Tarrio *et al.*, 2000; Cameron *et al.*, 2004; Ware *et al.*, 2008). For example, using the Mollusca + Annelida as an outgroup of Arthropoda, Nardi *et al.* (2003) challenged the traditional concept of Hexapoda forming a monophyletic clade. However, when different taxa were used as

Fund items: The Third Phase of the Innovation Program in the Chinese Academy of Sciences; Grants from the National Natural Science Foundation of China (Grant Nos. 30670242 and 30500056); the Major Program from the Natural Science Foundation, Beijing (6081002).

Brief Introduction of the First Author: LUO A-Rong, PhD student, majoring in ecology, E-mail: luoar@ioz.ac.cn

* Corresponding author, Tel.: 010-64807085; E-mail: zhucd@ioz.ac.cn

Received: 2009-08-26; Accepted: 2010-01-06

outgroup, the phylogeny of Nardi *et al.* (2003) was not consistently obtained (Cameron *et al.*, 2004). In addition, the research of Ware *et al.* (2008) on the relationships for Dictyoptera showed that outgroup selection influenced tree topologies to the extent that unexpected placements were obtained for some taxa.

1.2 Strategies of outgroup selection

Outgroup selection usually conforms to several principles. Whenever possible, the outgroup should be outside of, but closely related to the ingroup, and preferably the sister group of the ingroup, and the outgroup should contain more than one taxon; these guidelines help to avoid erroneous phylogenetic signals that can result in "random outgroup effect" and long branch attraction (LBA) sometimes associated with root placement (Watrous and Wheeler, 1981; Maddison *et al.*, 1984; Wheeler, 1990; Tarrio *et al.*, 2000; Dalevi *et al.*, 2001; Graham *et al.*, 2002; Sanderson and Shaffer, 2002; Cameron *et al.*, 2004; Bergsten, 2005). Nevertheless, the use of a closely related sister group may not always lead to the correct hypothesis of phylogenetic relationships, especially when evolutionary rates are relatively high (Lyons-Weiler *et al.*, 1998; Sanderson and Shaffer, 2002), or there is only a single taxon in the sister group (Smith, 1994; Sanderson and Shaffer, 2002).

Among the previous studies investigating outgroup selection for tree reconstruction, Smith (1994) proposed three general strategies of outgroup selection specifically for rooting molecular trees: a single taxon outgroup, multiple taxa within a single sister group, and single taxa from successive sister groups. He pointed out that, compared with the other two strategies, the option of multiple taxa within a single sister group was theoretically better based on his viewpoints of tree balance and homoplasy. This suggestion was corroborated by his empirical test of the echnoids (the ingroup). However, this test nowadays seems to be unconvincing because of insufficient data (six taxa) and several methodological concerns (*e. g.*, no statistical test for nodal strength, and compounding factors affecting tree topologies). Our study followed these three strategies while investigating phylogenetic relationships within the family Halictidae. Because it is unlikely to encounter a relative rate speedup in the sister group (Sanderson and Shaffer, 2002), we just assumed this has not occurred.

1.3 Phylogenetic relationships of the Halictidae

The monophyletic family Halictidae (Hymenoptera: Apoidea) is a group of short-tongued (S-T) bees. It is the second largest family in Apoidea with more than 4 150 described species (http://pick14. pick. uga. edu/mp/20q? guide = Apoidea_species), of which some species are the commonest bees (Packer and Taylor, 1997; Alexander and Michener, 1995; Michener, 2000). Other than the genus *Apis* (Apidae), halictids dominate other bees in numbers of individuals in many temperate areas (Michener, 2000). The great diversity in social behavior has made them a model group for studying social evolution (Danforth *et al.*, 2008).

Generally, Halictidae contains four monophyletic subfamilies: Halictinae, Rophitinae, Nomiinae and Nomioidinae. Although this subfamily classification is not universally accepted, their phylogenetic relationships are supported by both morphological and molecular evidence (Pesenko, 1999; Michener, 2000; Danforth *et al.*, 2004; 2008). The generally accepted subfamily level relationships are as follows: (Rophitinae (Nomiinae (Nomioidinae + Halictinae))). At the family level, Halictidae, Colletidae and Andrenidae of S-T bees are putatively monophyletic, but Stenotritidae is resolved either branching off within or as the sister group of Colletidae (Michener, 2000; Danforth *et al.*, 2006, 2006b). The phylogenetic relationships among them are as follows: (Andrenidae (Halictidae (Colletidae + Stenotritidae))) (Danforth *et al.*, 2006a, 2006b). In the long-tongued bees composed of the Megachilidae and the Apidae, the Megachilidae is monophyletic and falls outside of the family Halictidae (Danforth *et al.*, 2006a, 2006b).

Herein, we investigate the effect of outgroup selection on the construction of trees using a molecular phylogenetic analysis of Halictidae. Our outgroup selection strategies involve the sister groups of Halictidae as follows: Strategy Ⅰ, a single taxon in the sister group, but one that differs from Smith's (1994) single taxon outgroup; Strategy Ⅱ, multiple taxa within the sister group, which is similar to Smith's grouping; Strategy Ⅲ, multiple taxa from successive sister groups, different from his use of a single taxon from successive sister groups. Tree topologies derived from datasets with different strategies of outgroup selection would serve to evaluate the effects of outgroup selection.

2 METHODS AND MATERIALS

2.1 Gene marker

The gene 28S rDNA D2-D3 was used as the molecular marker because of its large proportion of potentially phylogenetically informative characters (Hancock and Dover, 1988; Hancock *et al.*, 1988; Tautz *et al.*, 1988) and widespread implementation to infer higher level relationships (De Rijk *et al.*, 1995; Schnare *et al.*, 1996; Danforth *et al.*, 2006a, 2006b). A moderate number of D2-D3 sequences from the family Halictidae and related taxa were taken from GenBank (http://www. ncbi. nlm. nih. gov/).

2.2 Taxa and datasets

Table 1 lists the taxa and GenBank accession numbers. For the family Halictidae, all 24 D2-D3 sequences available in GenBank (till 6/27/2008) were used as the ingroup. This sampling included its four subfamilies: Halictinae (6), Nomiinae (8), Nomioidinae (1) and Rophitinae (9). The outgroup included four taxa representing three of the four subfamilies of Andrenidae, seven taxa representing seven subfamilies of Colletidae, and three taxa representing two subfamilies of Megachilidae. These outgroup taxa were all chosen randomly to represent their taxonomic group. Besides, there was only one taxon from GenBank to represent the family Stenotritidae. All other apoids other than halictids can be used as potential outgroup members to Halictidae.

Table 1 Taxonomic diversity and nucleotide sequences obtained from GenBank used to investigate the effect of outgroup selection

Families	Subfamilies	Species	GenBank accession no.	Taxon no.
Halictidae	Halictinae	*Agapostemon tyleri*	AY654506	IG 1
	Halictinae	*Augochlorella pomoniella*	AY654507	IG 2
	Halictinae	*Halictus rubicundus*	AY654510	IG 3
	Halictinae	*Sphecodes pecosensis*	DQ072154	IG 4
	Halictinae	*Sphecodes* sp. Spsp1055	DQ072155	IG 5
	Halictinae	*Patellapis* (*Zonalictus*) sp. BND-2006	DQ060870	IG 6
	Nomiinae	*Dieunomia heteropoda*	DQ072151	IG 7
	Nomiinae	*Dieunomia nevadensis*	AY654512	IG 8
	Nomiinae	*Dieunomia nevadensis*	DQ060852	IG 9
	Nomiinae	*Lipotriches patellifera*	DQ072146	IG 10
	Nomiinae	*Macronomia aureozonata*	DQ072149	IG 11
	Nomiinae	*Nomia tetrazonata*	DQ072152	IG 12
	Nomiinae	*Pseudapis obesula*	DQ060868	IG 13
	Nomiinae	*Pseudapis unidentata*	AY654514	IG 14
	Nomioidinae	*Nomioides facilis*	AY654511	IG 15
	Rophitinae	*Conanthalictus conanthi*	DQ072144	IG 16
	Rophitinae	*Conanthalictus wilmattae*	AY654508	IG 17
	Rophitinae	*Dufourea mulleri*	AY654509	IG 18
	Rophitinae	*Penapis penai*	AY654513	IG 19
	Rophitinae	*Rophites algirus*	AY654515	IG 20
	Rophitinae	*Rophites algirus*	DQ072159	IG 21
	Rophitinae	*Systropha curvicornis*	AY654516	IG 22
	Rophitinae	*Systropha glabriventris*	DQ072156	IG 23
	Rophitinae	*Xeralictus bicuspidariae*	AY654517	IG 24
Andrenidae	Andreninae	*Andrena nasonii*	DQ060849	OG 25
	Oxaeinae	*Protoxaea gloriosa*	AY654480	OG 26
	Panurginae	*Calliopsis subalpinus*	DQ060850	OG 27
	Panurginae	*Protandrena nanulus*	DQ060857	OG 28
Colletidae	Colletinae	*Colletes graeffei*	EF363690	OG 29
	Diphaglossinae	*Caupolicana vestita*	AY654486	OG 30
	Euryglossinae	*Euryglossina globuliceps*	AY654490	OG 31
	Hylaeinae	*Hylaeus proximus*	AY654493	OG 32
	Paracolletinae	*Leioproctus irroratus*	AY654495	OG 33
	Scraptrinae	*Scrapter niger*	AY654501	OG 34
	Xeromelissinae	*Chilimelissa rozeni*	AY654481	OG 35
Megachilidae	Fideliinae	*Fidelia major*	AY654539	OG 36
	Megachilinae	*Lithurgus apicalis*	DQ072145	OG 37
	Megachilinae	*Megachile pugnata*	AY654543	OG 38
Stenotritidae	–	*Stenotritus* sp.	AY654503	OG 39

IG denotes ingroup taxa, while OG denotes outgroup taxa.

Four datasets were generated using the three strategies of outgroup selection (Table 2). Considering the family level relationships within the superfamily Apoidea, Stenotritidae (dataset I, 25 taxa) was selected to be the outgroup of Strategy I. Strategy II used Colletidae + Stenotritidae (dataset II, 32 taxa) as the outgroup. Strategy III contained Andrenidae + Colletidae + Stenotritidae (dataset III-a, 36 taxa) and Andrenidae + Colletidae + Megachilidae + Stenotritidae (dataset III-b, 39 taxa) as the outgroup. We also used seven extended datasets, in each of which one species from the seven taxa (as above) of Colletidae was used as the outgroup.

Table 2 Four standard datasets and extended datasets in the study of the effect of outgroup selection

	Standard datasets				Extended datasets				
Code	I	II	III-a	III-b	I-a	I-b	I-c	…	I-g
Taxa	IG + (39)	I + (29 – 35)	II + (25 – 28)	III-a + (36 – 38)	IG + (29)	IG + (30)	IG + (31)	…	IG + (35)

I, II, and III correspond to the three strategies of outgroup selection. Numbers in parentheses correspond to outgroup taxon numbers in Table 1.

2.3 Sequence alignments

All sequences (Table 1) were prepared for multiple sequence alignment. To avoid an alignment being dependent on the removed taxon sequence (Cameron *et al.*, 2004), 24 sequences of the Halictidae in combination with sequences of outgroup taxa/taxon were independently aligned using ClustalW (Thompson *et al.*, 1994) with default parameters. A final adjustment was made by eye. The end regions were truncated to ensure that only the D2-D3 region was used, with reference to the 28S rDNA sequence from the honey bee, *Apis mellifera* (Gillespie *et al.*, 2006). Alignment statistics were evaluated using PAUP* v.4.0b10 (Swofford, 2002).

2.4 Tree reconstruction

Phylogenetic analysis was performed with both PAUP* v.4.0b10 (Swofford, 2002) and MrBayes v. 3.1.2 (Huelsenbeck and Ronquist, 2001) for each of the standard datasets (I, II, III-a, III-b). For the extended datasets, Bayesian inference was not feasible because of long computation times.

Maximum likelihood Modeltest v.3.7 (Posada and Crandall, 1998; Posada and Buckley, 2004) was first employed to select the appropriate DNA substitution model for our aligned NEXUS files. After the favored model (TVM + I + G) and certain parameter values as a block appended to the original NEXUS files, the analyses were performed in PAUP* v. 4. 0b10 (Swofford, 2002). They were analyzed using heuristic analysis under the criterion of likelihood with 100 random addition sequence replications and tree bisection reconnection (TBR) branch swapping. We generally got one maximum likelihood (ML) tree for each dataset.

Maximum parsimony Trees were generated by PAUP* v. 4. 0b10 (Swofford, 2002). The heuristic search involved 2 000 random addition sequence replications under tree bisection reconnection (TBR) branch swapping. A 50% majority rule consensus tree was used for comparison (MP tree).

Bayesian inference First, MrModeltest v. 2. 2 (Nylander, 2004; Posada and Buckley, 2004) was specifically employed to select the appropriate DNA substitution model for Bayesian inference (BI). Two independent Markov Chain Monte Carlo (MCMC) analyses of 10 million iterations were performed in MrBayes v. 3. 1. 2, each with 4 chains, three hot, one cold, sampling one tree per 100 iterations (Huelsenbeck and Ronquist, 2001). The "sump" command together with "burnin = 25 000 (25% of the samples)" was used to determine the appropriate "burnin". If the parameters summarized showed that the potential scaled reduction factor (PSRF) was reasonably close to 1.0, "sumt burnin = 25 000" was then used to pool trees after discarding the "burnin".

2.5 Statistical analysis

Nonparametric bootstrapping was employed to estimate branch support of the MP and the ML trees. For the MP trees, we used 1 000 replications. For the ML trees, due to the huge computation times required, we used 500 replications. BI Posterior probabilities were used to estimate the reliability of the BI tree topologies.

3 RESULTS

3.1 Sequence alignments

The gene sequences of 28S rDNA D2-D3 contained 656 nucleotide sites, corresponding to the 28S sequence of the honey bee, *Apis mellifera*, from the end of Helix 531 to the beginning of Helix 589'. After alignments by ClustalW, both the number of total sites and the number of potentially parsimony informative sites in each of the four standard datasets differed (Table 3).

Table 3 Alignment statistics for four standard datasets

Datasets	Constant sites	Variable but uninformative sites	Parsimony-informative sites	Total sites
I	443 (62.75%)	103 (14.59%)	160 (22.66%)	706
II	406 (56.31%)	109 (15.12%)	206 (28.57%)	721
III-a	368 (51.18%)	114 (15.86%)	237 (32.96%)	719
III-b	350 (48.21%)	108 (14.88%)	268 (36.91%)	726

3.2 Tree topologies

Dataset I Monophyly of the subfamilies, except Nomioidinae, was supported by ML and BI analyses (Fig. 1). With MP, *Augochlorella pomoniella* clustered with Nomiinae rather than Halictinae; this arrangement did not support monophyly of either Halictinae or Nomiinae (Table 4).

Table 4 Summary of tree topologies for four standard datasets from three tree reconstruction methods

Dataset + Method	Nomiinae	Halictinae	Rophitinae	Nomiinae + Halictinae	Halictidae
I-ML	√	√	√	√	√
I-MP	×	×	√	√	√
I-BI	√	√	√	√	√
II-ML	√	√	√	√	√
II-MP	×	×	√	√	√
II-BI	√	√	√	√	√
III-a-ML	√	×	√	√	××
III-a-MP	√	×	√	√	××
III-a-BI	√	×	√	√	××
III-b-ML	√	√	√	√	√
III-b-MP	√	√	√	√	××
III-b-BI	√	×	√	√	√

ML: Maximum likelihood; MP: Maximum parsimony; BI: Bayesian inference. "√" denotes that the putative monophyly of a certain taxonomic group is corroborated by our results; "××" denotes that the monophyly is not obtained in our results due to the unexpected positions of some taxa; "×" denotes the uncorroborated monophyly of Nomiinae and Halictinae due to the position of *Augochlorella pomoniella*. The position of *Nomioides facilis* in tree topologies was neglected in this summary.

The MP and ML trees were evaluated for the extended datasets only. When *Colletes graeffei* was used as the outgroup, tree topologies were similar to those of dataset I. Using *Hylaeus proximus* as the outgroup, unresolved internal phylogenetic relationships were obtained within Halictidae. In the other trials, tree topologies were generally consistent with the accepted taxonomy except for the phylogenetic relationships of a few taxa (*e. g.*, *Nomioides facilis* and *Augochlorella pomoniella*).

Dataset II The ML and BI trees (Fig. 2), resolved the monophyly of three subfamilies and their current phylogenetic relationships. Again, the monophyly of Halictinae was not supported by MP tree because of the position of *Augochlorella pomoniella*.

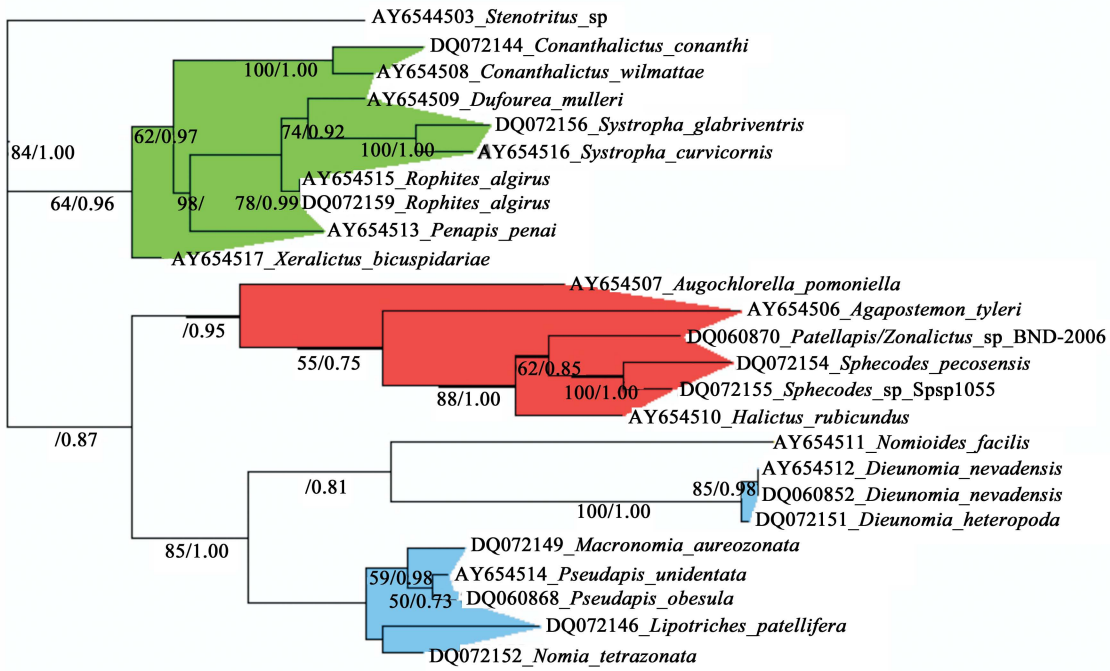

Fig. 1 Tree topology for halictid bees derived from maximum likelihood and Bayesian inference methods using DNA sequence data in the standard dataset I

Node labels are given as bootstrap values/Bayesian posterior probabilities. Values contrary to the majority rule are not shown. Blue background represents the taxa in Nomiinae; red, the Halictinae; green, the Rophitinae; gray, the outgroup; and one yellow leaf, the Nomioidinae.

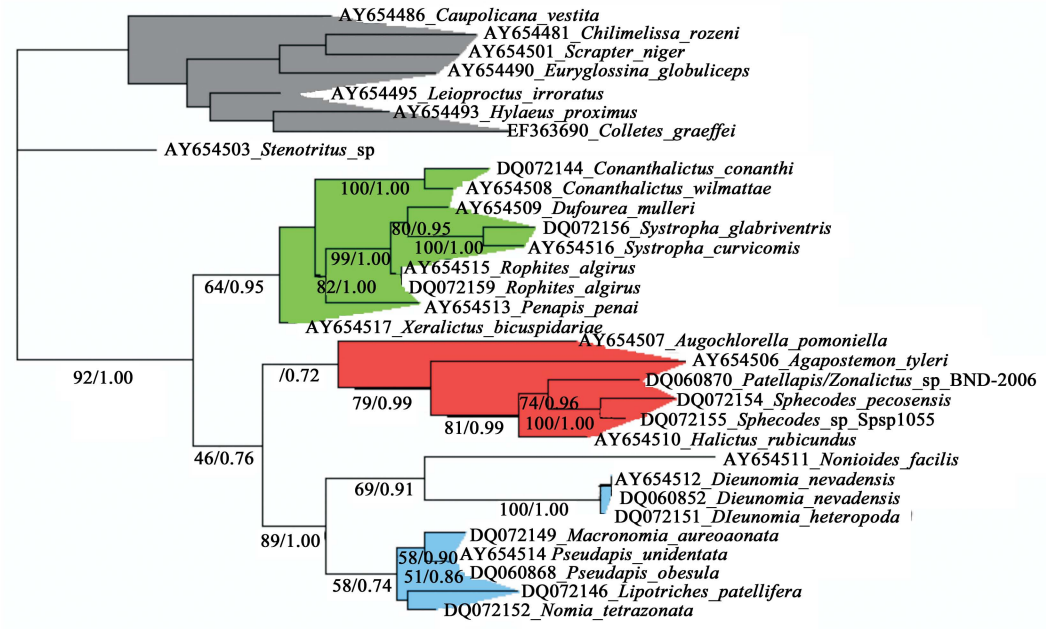

Fig. 2 Tree topology for halictid bees derived from maximum likelihood and Bayesian inference methods using DNA sequence data in standard dataset II

Node labels are given as bootstrap values/Bayesian posterior probabilities. Values contrary to the majority rule are not shown. Blue background represents the taxa in Nomiinae; red, Halictinae; green, Rophitinae; gray, the outgroup; and one yellow leaf, Nomioidinae.

Dataset III-a When four representatives of Andrenidae were added to the outgroup (dataset III-a), the monophyly of the family Halictidae was not resolved in any of the three analyses, although the monophyly of Rophitinae was always supported. In the ML and BI trees, four representatives of Andrenidae, part of the outgroup, clustered with Nomiinae + Halictinae + *Nomioides facilis* as their sister group (Fig. 3). However, MP tree united the representatives of Andrenidae with the subfamily Rophitinae in a paraphyletic clade at the base of Nomiinae + Halictinae + *Nomioides facilis*. Another common characteristic of these four trees was that *Augochlorella pomoniella* clustered at the base of Nomiinae + *Nomioides facilis*.

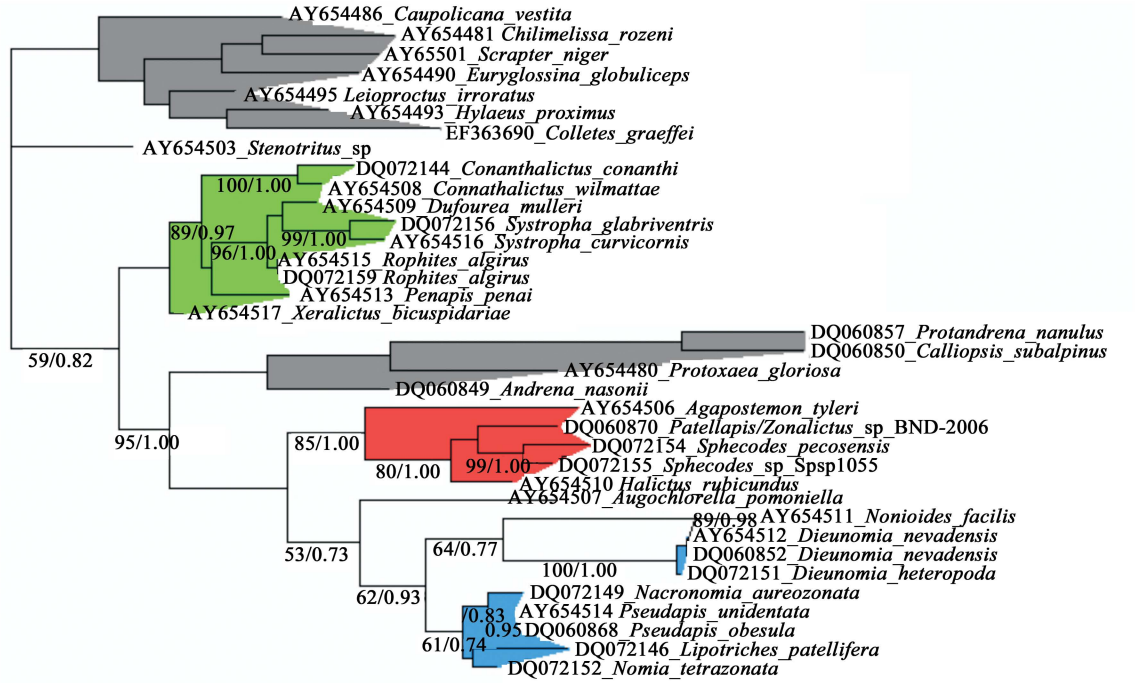

Fig. 3 Tree topology for halictid bees derived from maximum likelihood and Bayesian inference methods using DNA sequence data in standard dataset III-a

Node labels are given as bootstrap values/Bayesian posterior probabilities. Values contrary to the majority rule are not shown. Blue background represents the taxa in Nomiinae; red, the Halictinae; green, the Rophitinae; gray, the outgroup; and one yellow leaf, the Nomioidinae.

Dataset III-b When the three representatives of Megachilidae were added to the outgroup (dataset III-b), the monophyly of the family Halictidae was recovered (Fig. 4; Fig. 5) in all trees except for MP tree, in which four representatives of Andrenidae together with the three samples of Megachilidae clustered as the sister group of Nomiinae + Halictinae + *Nomioides facilis*.

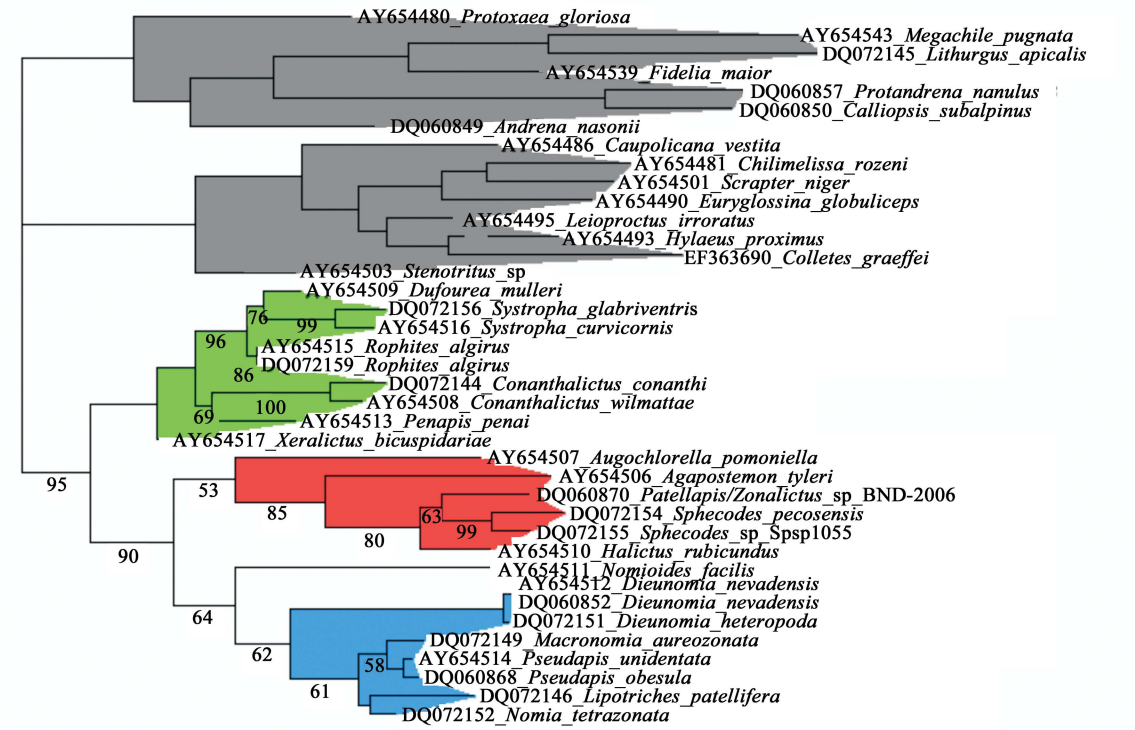

Fig. 4 Tree topology for halictid bees derived from maximum likelihood using DNA sequence data in standard dataset III-b
Node labels are bootstrap values. Blue background represents the taxa in Nomiinae; red, the Halictinae; green, the Rophitinae; gray, the outgroup; and one yellow leaf, the Nomioidinae.

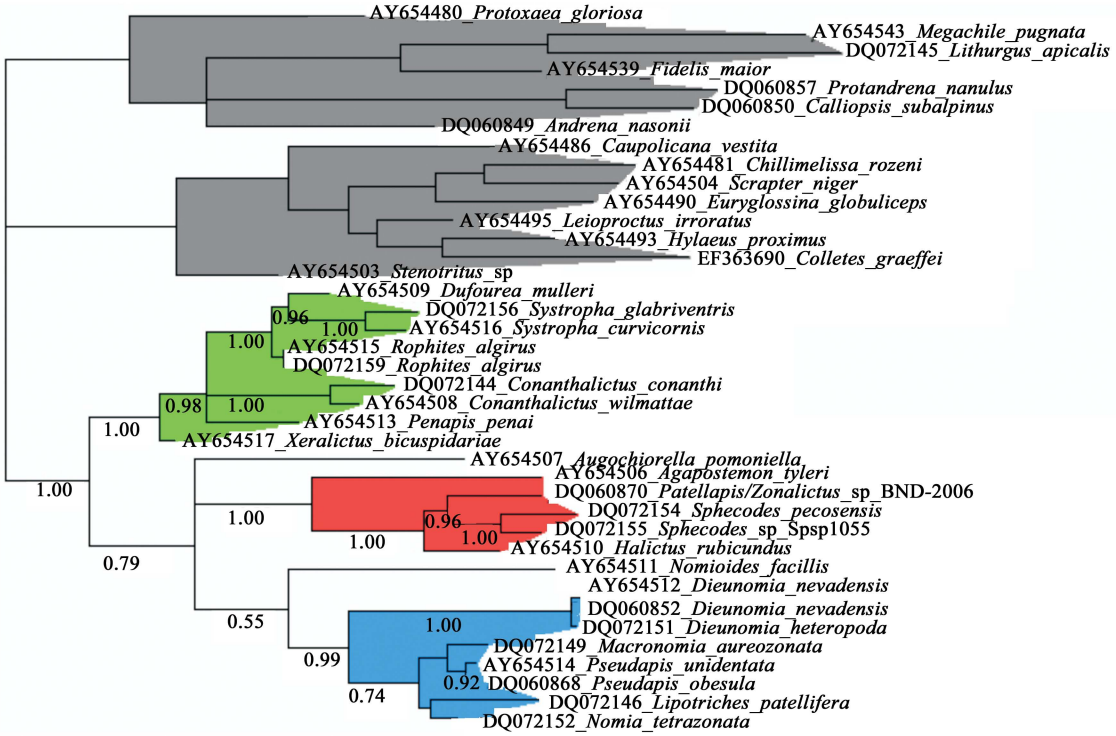

Fig. 5 Tree topology for halictid bees derived from Bayesian inference using DNA sequence data in standard dataset III-b
Noda labels are the Bayesian posterior probabilities. Blue background represents the taxa in Nomiinae; red, the Halictinae; green, the Rophitinae; gray, the outgroup; and one yellow leaf, the Nomioidinae.

4 DISCUSSION

We selected the family Halictidae as the ingroup to evaluate the effect of outgroup selection on tree reconstruction. One gene marker (28S rDNA D2-D3) was employed because of its established phylogenetic signal and established resolving power within higher levels of Halictidae. We sought to remove many factors that could affect tree topology by controlling variables such as sequence region and number of tree generations. Then, differences in tree topologies should have reflected differences in outgroup composition and method of tree reconstruction. Nevertheless, the effect of certain gene marker (*i. e.*, 28S rDNA D2-D3 in our case) cannot be removed, and further research on other genes still needs to be done in the future.

Our evaluation assumed that the putative topology of Halictidae by Danforth *et al.* was correct due to sufficient taxon sampling, the incorporation of three gene markers and the application of powerful methods of analysis (Danforth *et al.*, 2004, 2008). *Nomioides facilis* was seldom considered in the assessment of accuracy because its usual grouping with genus *Dieunomia* may have owed to long branch attraction (Figs. 1 – 3), and/or inadequate taxon sampling. Unfortunately, only one sequence was available for Nomioidinae. Without considering the *Nomioides facilis*, some trees we got supported the putative phylogeny, but others did not.

Our analysis demonstrated that multiple taxa within the nearest sister group should be used when forming outgroups. Trees rooted with a greater diversity of immediate sister group species had topologies most congruent with the putative phylogeny. This conclusion is tempered by the discovery that multiple taxa from successive sister groups resulted in very unstable tree topologies, even when multiple taxa of the sister group of Halictidae were included in the outgroup. The effect was so great that more distant sister group (Andrenidae) clustered with one subclade of the ingroup, and thus rearranged the ingroup tree topology greatly. When we adopted Strategy III for outgroup formation, we obtained an untenable topology in the ingroup. Conversely, application of Strategy I on the standard dataset together with the extended datasets revealed that a single taxon in sister group could represent the outgroup, if the selected taxon accurately polarized characters in the ingroup, and depending on the method of tree reconstruction. This discovery suggested that phylogenies may be consistently reconstructed even when the outgroup contains few taxa, as long as the nearest sister group taxon is used.

Synapomorphies are the fundamental basis of phylogeny reconstruction. Unlike many classes of morphological characters, nucleotide sequences only have five possible character states, including indels. Outgroup choice should be critical because homoplasy owing to parallel change could be a troublesome factor in tree reconstruction. In our case, Andrenidae was used as the most distantly related outgroup member to Halictidae. This arrangement affectively added outgroup taxa that fell outside of the ingroup-plus-first-outgroup (Sanderson and Shaffer, 2002). Did parallel nucleotide substitutions result in homoplasy at some nucleotide positions to the extent of causing changes in the tree? When Andrenidae was used as the outgroup in dataset III-a, key nucleotide sites had homoplastic signals that broke the monophyly of Halictidae. However, when three taxa in Megachilidae were added, much of this confusion was eliminated because the Megachilidae taxa avoided homoplastic signals to influence the tree reconstruction; the monophyly of Halictidae was recovered again. For single taxon outgroups, even the nearest sister group, homoplasy should exist (Smith, 1994), but the extent of homoplasy and the attainment of false phylogenetic relationships depends on the taxon itself. Depending on the outgroup taxon, some tree topologies were inconsistent with the putative tree (*e. g.*, extended dataset I-e), yet others were largely congruent (*e. g.*, dataset I). However, multiple taxa within the sister group could shelter homoplasy coming from certain taxon, and the whole of them tend to result in correct phylogeny.

Three different methods of tree reconstruction were employed to evaluate the consistency of tree topology for each standard dataset. Although the assumptions differ from the optimality criteria to the extent of some having advantages (*e. g.*, realism, generality, and economy of assumptions) over others (Goloboff, 2003), MP seemed to consistently provide less resolution than ML and BI approaches. However, MP produced more consistent results depending on outgroup strategy. Thus, outgroup choice may be particularly critical for model-based approaches to tree reconstruction, in particular ML and BI, at least in this case study. Further trials are required to determine whether this finding is a generality, or not.

ACKNOWLEDGEMENTS We thank those who generated, edited, and submitted the related DNA sequences. Also, we thank Dr. J. S. Noyes in the Natural History Museum (London, UK) and Prof. A. P. Vogler in the Imperial College (London, UK), who provided facilities for large data mining of DNA sequences when C. D. Zhu visited both institutions from 2004 until early 2006.

References

Alexander BA, Michener CD, 1995. Phylogenetic studies of the families of short-tongued bees. *Univ. Kansas Sci. Bull.*, 55: 377 – 424.

Bergsten J, 2005. A review of long-branch attraction. *Cladistics*, 21: 163 – 193.

Cameron SL, Miller KB, D'Haese CA, Whiting MF, Barker SC, 2004. Mitochondrial genome data alone are not enough to unambiguously resolve the relationships of Entognatha, Insecta and Crustacea *sensu lato* (Arthropoda). *Cladistics*, 20: 534 – 557.

Cavalli-Sforza LL, Edwards AW, 1967. Phylogenetic analysis: Models and estimation procedures. *Am. J. Hum. Genet.*, 19: 233 – 257.

Dalevi D, Hugenholtz P, Blackall LL, 2001. A multiple-outgroup approach to resolving division-level phylogenetic relationships using 16S rDNA data. *Int. J. Syst. Evol. Microbiol.*, 51: 385 – 391.

Danforth BN, Brady SG, Sipes SD, Pearson A, 2004. Single copy nuclear genes recover Cretaceous-age divergences in bees. *Syst. Biol.*, 53: 309 – 326.

Danforth BN, Eardley C, Packer L, Walker K, Pauly A, Randrianambinintsoa FJ, 2008. Phylogeny of Halictidae with an emphasis on endemic African Halictinae. *Apidologie*, 39: 86 – 101.

Danforth BN, Fang J, Sipes S, 2006a. Analysis of family-level relationships in bees (Hymenoptera: Apiformes) using 28S and two previously unexplored nuclear genes: CAD and RNA polymerase II. *Mol. Phylogenet. Evol.*, 39: 358 – 372.

Danforth BN, Sipes S, Fang J, Brad SG, 2006b. The history of early bee diversification based on five genes plus morphology. *Proc. Natl. Acad. Sci. USA*, 103: 15 118 – 15 123.

De Rijk P, Van de Peer Y, Chapelle S, De Wachter R, 1994. Database on the structure of large ribosomal subunit RNA. *Nucleic Acids Res.*, 22: 3 495 – 3 501.

Gillespie JJ, Johnston JS, Cannone JJ, Gutell RR, 2006. Characteristics of the nuclear (18S, 5.8S, 28S and 5S) and mitochondrial (12S and 16S) rRNA genes of *Apis mellifera* (Insecta: Hymenoptera): Structure, organization, and retrotransposable elements. *Insect Mol. Biol.*, 15: 657 – 686.

Goloboff PA, 2003. Parsimony, likelihood and simplicity. *Cladistics*, 19: 91 – 103.

Graham SW, Olmstead RG, Barrett SCH, 2002. Rooting phylogenetic trees with distant outgroups: a case study from the commelinoid monocots. *Mol. Biol. Evol.*, 19: 1 769 – 1 781.

Hancock JM, Dover GA, 1988. Molecular coevolution among cryptically simple expansion segments of eukaryotic 26S/28S rRNAs. *Mol. Biol. Evol.*, 5: 377 – 392.

Hancock JM, Tautz D. Dover GA, 1988. Evolution of the secondary structures and compensatory mutations of the ribosomal RNAs of *Drosophila melanogaster*. *Mol. Phylogenet. Evol.*, 5: 393 – 414.

Huelsenbeck JP, Ronquist FR, 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17: 754 – 755.

Lyons-Weiler J, Hoelzer GA, Tausch RJ, 1998. Optimal outgroup analysis. *Biol. J. Linn. Soc.*, 64: 493 – 511.

Maddison WP, Donoghue MJ, Maddison DR, 1984. Outgroup analysis and parsimony. *Syst. Zool.*, 33: 83 – 103.

Michener CD, 2000. The Bees of the World. The Johns Hopkins University Press, Baltimore and London. 304-338.

Milinkovitch MC, Lyons-Weiler J, 1998. Finding optimal ingroup topologies and convexities when the choice of outgroups is not obvious. *Mol. Phylogenet. Evol.*, 9: 348 – 357.

Nardi F, Spinsanti G, Boore JL, Carapelli A, Dallai R, Frati F, 2003. Hexapod origins: Monophyletic or paraphyletic? *Science*, 299: 1 887 – 1 889.

Nylander JAA, 2004. MrModeltest v2, program distributed by the author. Evolutionary Biology Centre, Uppsala University. Available from: http://www. abc. se/~ nylander/mrmodeltest2/mrmodeltest2. html.

Packer L, Taylor JS, 1997. How many hidden species are there? An application of the phylogenetic species concept to genetic data for some comparatively well known bee "species". *Can. Entomol.*, 129: 587 – 594.

Pesenko YA, 1999. Phylogeny and classification of the family Halictidae revised (Hymenoptera: Apoidea). *J. Kansas Entomol. Soc.*, 72: 104 – 123.

Phillips A, Janies D, Wheeler W, 2000. Multiple sequence alignment in phylogenetic analysis. *Mol. Phylogenet. Evol.*, 16: 317 – 330.

Posada D, Buckley TR, 2004. Model selection and model averaging in phylogenetics: Advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Syst. Biol.*, 53: 793 – 808.

Posada D, Crandall KA, 1998. ModelTest: Testing the best-fit model of nucleotide substitution. *Bioinformatics* 14: 817 – 818.

Sanderson MJ, Shaffer HB, 2002. Troubleshooting molecular phylogenetic analyses. *Annu. Rev. Ecol. Syst.*, 33: 49 – 72.

Schnare MN, Damberger SH, Gray MW, Gutell RR, 1996. Comprehensive comparison of structural characteristics in eukaryotic cytoplasmic large subunit (23S-like) ribosomal RNA. *J. Mol. Biol.*, 256: 701 – 719.

Smith AB, 1994. Rooting molecular trees: Problems and strategies. *Biol. J. Linn. Soc.*, 51: 279 – 292.

Swofford DL, 2002. PAUP*: Phylogenetic Analysis Using Parsimony (* and Other Methods), Version 4b 10. Sinauer, Sunderland, Massachusetts.

Tarrio R, Rodriguez-Trelles F, Ayala F, 2000. Tree rooting with outgroups when they differ in their nucleotide composition from the ingroup: the *Drosophila saltans* and *willistoni* groups, a case study. *Mol. Phylogenet. Evol.*, 16: 344 – 349.

Tautz DJ, Hancock JM, Webb DA, Tautz C. Dover GA, 1988. Complete sequences of the rRNA genes of *Drosophila melanogaster*. *Mol. Biol. Evol.*, 5: 366 – 376.

Thompson JD, Higgins DG, Gibson TJ, 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22: 4 673 – 4 680.

Ware JL, Litman J, Klass KD, Spearman LA, 2008. Relationships among the major lineages Dictyoptera: the effect of outgroup selection on dictyopteran tree topology. *Syst. Entomol.*, 33: 429 – 450.

Watrous LE, Wheeler QD, 1981. The out-group comparison method of character analysis. *Syst. Zool.*, 30: 1 – 11.

Wheeler WC, 1990. Nucleic acid sequence phylogeny and random outgroups. *Cladistics*, 6: 363 – 367.

外群选择对隧蜂科(膜翅目:蜜蜂总科)系统重建的影响

罗阿蓉[1,2],张彦周[1],乔慧捷[1],史卫峰[3],Robert W. MURPHY[4], 朱朝东[1,*]

(1. 中国科学院动物研究所动物系统与进化重点实验室,北京 100101; 2. 中国科学院研究生院,北京 100049;
3. UCD Conway Institute of Biomolecular and Biomedical Sciences, University College Dublin, Dublin 4, Ireland;
4. Department of Natural History, Royal Ontario Museum, 100 Queen's Park, Toronto, ON, M5S 2C6, Canada)

摘要: 外群用于给树附根和推断祖先性状状态。通常,来自内群的姐妹群中的多个分类单元被共同选择作为外群。为了在经验上验证这一方法,我们采用了 3 种外群选择策略:姐妹群中的单一分类单元,姐妹群中的多个分类单元和连续姐妹群中的多个分类单元。以隧蜂科(膜翅目:蜜蜂总科)的系统发育重建为例,我们评估了这 3 种策略对树拓扑结构的影响,包括最大似然树、最大简约树和贝叶斯树。初步结果表明:相比其他两种策略,采用姐妹群中的多个分类单元作为外群更有利于系统发育重建得到现已被广泛认可的隧蜂科系统发育关系;相比最大似然法和贝叶斯法,虽然隧蜂科系统发育关系没有被很好地解决,但最大简约法在不同外群选择策略下得到了较为一致的拓扑结构。

关键词: 隧蜂科; 非同源相似; 单系性; 外群; 姐妹群; 系统发育

中图分类号: Q969　　**文献标识码**: A　　**文章编号**: 0454-6296(2010)02-0192-10

(责任编辑: 袁德成)